Gaze+RST: Integrating Gaze and Multitouch for Remote Rotate-Scale-Translate Tasks

Jayson Turner[1] **Jason Alexander**[1] **Andreas Bulling**[2] **Hans Gellersen**[1]

[1]Lancaster University
Lancaster, United Kingdom
{j.turner, j.alexander, h.gellersen}@lancaster.ac.uk

[2]Max Planck Institute for Informatics
Saarbrüken, Germany
andreas.bulling@acm.org

ABSTRACT

Our work investigates the use of gaze and multitouch to fluidly perform rotate-scale-translate (RST) tasks on large displays. The work specifically aims to understand if gaze can provide benefit in such a task, how task complexity affects performance, and how gaze and multitouch can be combined to create an integral input structure suited to the task of RST. We present four techniques that individually strike a different balance between gaze-based and touch-based translation while maintaining concurrent rotation and scaling operations. A 16 participant empirical evaluation revealed that three of our four techniques present viable options for this scenario, and that larger distances and rotation/scaling operations can significantly affect a gaze-based translation configuration. Furthermore we uncover new insights regarding multimodal integrality, finding that gaze and touch can be combined into configurations that pertain to integral or separable input structures.

Author Keywords

eye-based interaction; manipulation; cross-device; touch

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces—input devices and strategies

INTRODUCTION

This paper investigates the design and evaluation of gaze-supported interaction techniques that allow the fundamental tasks of rotate, scale, and translate (RST) on large remote displays. Proximity can be an inhibiting factor in large displays as users are unable to reach all areas from a single stationary location or height. Users need to be able to select, reposition, and manipulate content. These operations aid the practicality of large displays for collaboration, exploring large data sets [12], and sharing content [4, 24].

To interact with remote displays of this nature we propose the use of indirect multitouch pan, pinch, and rotation gestures, where content resides on the large display, and input is performed through a hand-held tablet device. This is motivated by the ubiquity of multitouch mobile devices. Such a configuration allows for compound input to perform RST tasks. However to be successful, a mapping is required to link input to specific locations of a large display.

In this work we propose the use of gaze to enable interaction between a user and a display. Our eyes look where we wish to interact, and this natural behaviour affords a quick and implicit pointing mechanism [26, 17]. Combined with touch on a mobile device, gaze has shown success in remote content positioning [18], and cross-device transfer [19]. In the techniques we present in this paper, gaze denotes the relative origin for multitouch input. We expand upon a gaze-supported manipulation concept reported by Stellmach and Dachselt [18]. They showed a technique with separate modes for positioning and manipulation. In our investigation we look at how users can orchestrate gaze and multiple degrees-of-freedom (DOF) concurrently, for one seamless interaction.

We use multitouch pan, pinch, and rotation gestures due to their expressive power and popularity. These gestures allow three dimensions–rotation, scaling, and translation–to be controlled with one hand while leaving the other hand free to hold the input device. Multitouch RST works well due to the *integral* nature of input and task. Dimensions can be controlled concurrently without restriction, and single hand input forms a unitary whole, perceptually relating all operations [11]. Conversely, a task involving translation and colour change would require a *separable* input structure, either using a single device with a mode switch, or two separate modalities [8]. Part of our work examines if gaze and touch, a multimodal input combination, can exhibit integrality.

There are three questions we aim to tackle through technique design and empirical evaluation (1) *Can gaze offer any performance benefits during remote RST tasks?* As gaze is fast, and naturally indicates the intended location of interaction, can we improve performance over touch-only RST. (2) *Do translation distance, rotation, and scale factors affect performance?* How do different configurations of RST cope with varying task factors, can gaze balance out the cost of larger distances and other operations? (3) *Do different configurations of gaze and RST affect integrality?* As we use a multimodal approach, can gaze can be blended with multitouch while maintaining similar integrality to touch alone?

We have designed four techniques that aim to address these questions. In our design process we limited the role of gaze to translation as gaze and translation operate in the same dimension. Each technique strikes a different balance between

CHI 2015, April 18 - 23 2015, Seoul, Republic of Korea
Copyright 2015 ACM 978-1-4503-3145-6/15/04...$15.00
http://dx.doi.org/10.1145/2702123.2702355

gaze- and touch-based influence over translation, and this balance is informed by different predicted levels of integrality.

We summarise our techniques here:

(1) Touch Translate (**TouchT**), this technique acts a baseline, using gaze only to point at objects for remote selection. All translation, scaling, and rotation actions are performed by pan, pinch, and rotation touch gestures. This is a classic integral input structure.

(2) Gaze Translate (**GazeT**), this is what we believe to be the most typical approach to combining gaze and RST. Gaze is used to point to and translate objects, while concurrent touch pinch and rotation gestures perform manipulation. If we treat gaze and touch as a unitary whole then the input structure is integral. However, if classed as separate modalities, gaze translation can be performed without influencing other control dimensions, and therefore input becomes separable.

(3) MAGIC Translate (**MagicT**), inspired by Zhai et al. [26], gaze and touch are cascaded, all translation, rotation and scaling are controlled by touch, however as the user translates, the object is snapped to the location of gaze, effectively speeding up translation. This technique aims to maintain the 'feeling' of integrality while using gaze to increase translation speed without separating primary translation control from touch.

(4) Gaze-Guided Touch Translate (**GazeGuidedTouchT**), touch performs translation, rotation, and scaling but the object is tied to a fixed linear path between gaze and the object location. The aim is to aid targeting while keeping the input structure fully integral.

The contribution of this work is two-fold. First we contribute the design and rationale of four techniques that enable gaze-supported remote RST on large displays. Second, an empirical evaluation with 16 participants has gained quantitative and qualitative insights that highlight the advantages and disadvantages of our designs. Gaze did not improve performance in remote RST tasks. We learn that gaze and multitouch can be configured in integral or separable input structures, leading to different performance effects over rotation and scaling factors. Furthermore we learn that in order to achieve integrality, gaze must be secondary or cascaded in configuration. Using gaze as the primary positioning modality causes difficulty in orchestrating multiple degrees-of-freedom.

RELATED WORK

Large Display Interaction

A variety of methods have been proposed for interaction with large remote displays. Keefe et al. have investigated interaction with large data visualisations for collaboration [12]. The work re-enforces our motivation, highlighting the need for fluid interaction with very large displays. Their system used mobile multitouch and proximity to perform input. Similarly Dachselt and Buchholz demonstrated the use of a mobile phone to throw content onto remote displays, tilting the mobile device to perform remote interaction [4]. An alternative approach by Boring et al. used the camera view of a smartphone to project touches onto displays that are in view [3].

Input based on hand-movement has shown success in large display environments. Han et al. used two sensors, one held in each hand to perform selection, translation and manipulation tasks in 3D interaction [9]. Each sensor was aware of its own attitude and position in space, enabling a number of gestures. In combination with hand-held devices, mid-air gestures can afford additional dimensions of control in large display input. Nancel et al. developed 12 techniques that combined in-air pointing, hardware-buttons, touch gestures, and mid-air gestures for remote pan-and-zoom [15]. Interestingly, touch-based input was found to have less fatigue than mid-air technique combinations. Work by Vogel et al. considered the importance of transitions between remote and up-close interaction. Using 3D glove input they compared ray-casting to relative input, finding that ray-casting reduced clutching [22].

Gaze-enhanced Selection, Positioning & Manipulation

Gaze-supported selection has been investigated extensively, showing that: gaze input can be made reliable with manual confirmation [10]; that gaze acts as predictor for input location [26]; and that the eyes can be faster for pointing than traditional mouse input [17, 6]. Yoo et al. combined gaze and mid-air gestures [24], using the eyes to act as an origin for pan-and-zoom on a large display.

Work by Stellmach and Dachselt is most relevant to our aims [18]. Their *Touch-enhanced Gaze Pointer* (TouchGP) technique used gaze for fine- and coarse-grained selection and positioning of objects, using touch on a smartphone to delimit actions. We have previously developed a similar technique, but it did not support high fidelity positioning [20]. TouchGP, contains two explicit modes: (1) remote positioning, and (2) manipulation control. By rotating the smartphone, users could disable positioning and instead use two thumbs or tilt to rotate and scale objects in place. The work presented no quantitative data on manipulation, although suggestions were made around seamlessly integrating manipulation and positioning.

Pfeuffer et al. presented an interaction concept combining gaze and touch within a single display [16]. They used gaze to select, move, and manipulate objects, redirecting touch input to distant locations on the display. Their work differs from Stellmach and Dachselt, and our own as gaze is not used for positioning. Our techniques take a similar approach regarding rotation and scaling, gaze denotes the object and multitouch transformations are applied through input redirection.

Complementary to our work, gaze can be used to create dynamic cursor sensitivity [5]. Fares et al. showed that the amount of gain applied to manual input can be dynamically altered depending on the location of gaze. We utilise this in our **TouchT** and **GazeGuidedTouchT** techniques to allow for high gain when crossing large distances, and low gain when performing final positioning.

Integrality & Separability

RST allows for the parallel manipulation of four DOF–*x*, *y*, *rotation*, and *scale*. RST can be classified as *integral* in accordance with principles evaluated by Jacob et al. [11]. The work states that an integral task allows users to cut across all

dimensions of control in a euclidian manner. Rotation, scaling, and translation form a unitary whole, as does the two-fingered multitouch input mechanism pan, pinch and rotate. Because of this perceptual matching between task structure and input, higher performance is likely. In contrast, a *separable* technique would only allow movement across one dimension at a time in a city-block fashion, but if matched to an appropriate task, would also show good performance. The work demonstrated that a mismatch between task and control structures can reduce performance. Integrality and separability are based on earlier work from Garner [7].

Contrary to the work above. It has been shown that users are not always able to manipulate all DOF simultaneously throughout a task. This is highlighted by Veit et al. stating that Jacob et al. did not take users' aptitude into consideration [21]. Users may decompose tasks into smaller sub-tasks that have less DOF, therefore integral tasks were performed in a separable manner. This is also true of physical objects, Wang et al. showed that transportation dominates orientation, making the two processes distinct [23]. Similar results were found by Martinet et al. [13] in multitouch 3D manipulation experiments, also with concerns over aptitude.

Less work has examined multimodal integrality and separability. Grasso et al. question the performance of mouse and speech input when selecting textual items [8]. The work found that unimodal input was best suited to integral tasks, whereas multimodal input was more appropriate for separable tasks. It is difficult to generalise this work to our own aims as speech does not operate in the same control space.

We are interested to understand if users can concurrently operate multiple DOF, and we feel that measuring integrality is suited to our goals. However there are alternative measures to quantify coordination. Zhai and Milgram regard coordinated movement as *efficient* movement [25]. Their method followed the logic that if the shortest trajectory is followed, then movement is efficient and thus coordinated. Balakrishnan and Hinckley explored symmetric bimanual interaction, regarding *parallel* input patterns as a positive outcome [1]. The work developed a new Parallelism measure, adapted from Masliah [14], to quantify the coordination they observed.

The hand follows the eyes, thus gaze is intrinsically tied to tasks through visual perception and motor control. The eyes are fast and reach a target sooner than manual input [26]. In light of these properties, we are interested to know how best to combine gaze and RST to create a usable technique, and whether gaze and touch comply with the above findings–that multimodal input is not suitable for integral tasks. It is difficult to generalise from the above work as speech does not operate in the same control space. The techniques we evaluate vary the role of gaze in an integral task.

REMOTE GAZE+RST TECHNIQUE DESIGN

Here we describe our four techniques and the rationale for our design choices. Each technique is designed for a large display-with-tablet context. Each are intended to be capable of translation over the full span of a large projected display, while allowing concurrent rotation and scaling. Each technique is different in its balance of gaze and manual input for translation, which we suspect will affect overall integrality. By varying the design of each technique in this way, we aim to learn which configuration is most appropriate for the task.

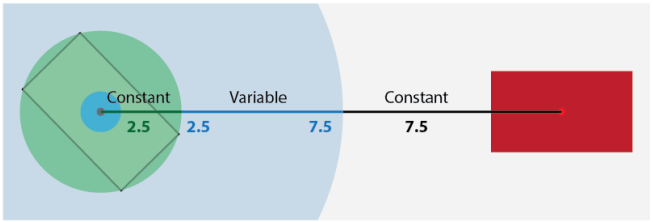

Figure 1. Gaze-contingent CDGain: Applied CDGain is dependent on object and gaze proximity (not to scale). Gain is constantly 2.5 at distances under $6\,^\circ$, linearly variable between 2.5 ($6\,^\circ$) and 7.5 ($10\,^\circ$), and a constant 7.5 over $10\,^\circ$ distance.

Control-Display Gain. Our **TouchT** and **GazeGuidedTouchT** techniques use a gaze-contingent control-display gain (CDGain) based on Magic Sense [5]. CDGain allows these techniques to cross large distances quickly while maintaining high-fidelity in final positioning, these techniques would otherwise be confounded. The amount of CDGain applied to touch-based translation is defined by the object's proximity to gaze. This strategy works as users look at the target destination pre-emptively, providing highest fidelity where attention is located. As shown in Figure 1, within close proximity ($6\,^\circ$) CDGain is set to 2.5. Between $6\,^\circ$ and $10\,^\circ$, CDGain is linearly variable between 2.5 and 7.5, this affects the speed of translation as the object approaches or moves away from gaze, easing the user into each extreme of CDGain. Beyond $10\,^\circ$, CDGain is set to 7.5. For **GazeT** and **MagicT**, CDGain is always 2.5 under multitouch translation, occurring only in close-proximity to gaze.

These CDGain values are based on the amount of tablet-based translation required to cover the width of the projected display. A CDGain of 2.5 allows the user to translate one full tablet display width to cover the full projected display ($2560/1024 = 2.5$), and 7.5 equates to one-third of the tablet width ($1024/3 = 341$, $2560/341 = 7.5$).

Technique Designs

Each of our techniques follow a common flow of interaction: (1) all objects are selected in the same way, users look at an object and hold down two fingers on the tablet device; (2) after selection, all techniques allow full touch control over an object (RST), provided the object and gaze remain close to each other (within a circular mask); (3) at this stage, objects are moveable between their selected origin and the target, and this is where each technique differs (explained below); (4) upon arrival at the target, the user can transition back to full manual control for final translation and manipulation; (5) throughout all stages of interaction users are able to perform rotation and scaling operations using multitouch.

The full touch control stages at the start and end of each technique are derived from Stellmach and Dachselt who showed that touch can combat imprecise gaze-based positioning [18].

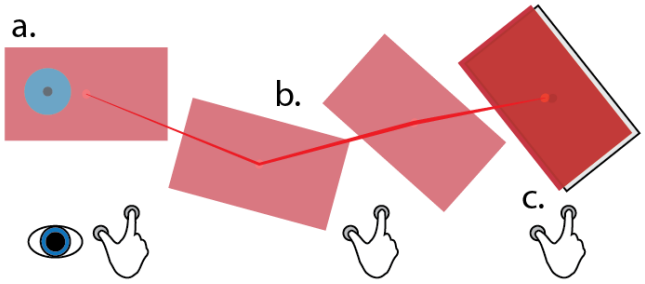

Figure 2. TouchT: (a) Look at object, touch on tablet to select. (b) Touch translates and manipulates object. (c) Position and drop.

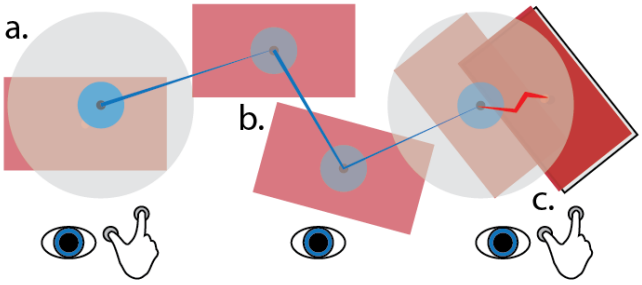

Figure 3. GazeT: (a) Look at object, touch on tablet to select, touch translates and manipulates while gaze is within displayed mask. (b) Gaze leaves mask, object follows gaze, touch manipulates. (c) Translation on tablet detaches object from gaze, touch performs final translation and manipulation.

As an example, the positioning of an object using gaze can be overridden by touch when fine-grained control is required.

TouchT

This is the most basic of our designs, all stages of the technique are controlled by touch, aside from initial gaze-pointing to highlight the object. Touching with two fingers on the tablet display confirms selection. The object is translated and manipulated using touch until it is dropped by releasing touch at a final destination (see Figure 2). This technique is an indirect version of typical tabletop or tablet RST, where the object and hand would reside in the same space. This technique uses the gaze-contingent CDGain described earlier to enable faster long-distance translation. As touch controls all dimensions throughout interaction, we would expect this technique to show integral behaviour.

GazeT

This technique is a modified version of Stellmach et al's Touch-enhanced Gaze Pointer (TouchGP) technique [18]. We extend this technique by allowing for rotation and scaling to take place during gaze-based positioning, whereas the original TouchGP technique used an explicit mode switch to change between positioning and manipulation modes.

In Figure 3(a) once an object is selected, a mask appears around the selection point. While gaze is within this mask, touch controls translation. The mask is contingent on the object–gaze and the object must remain close to allow full touch control. (b) Once gaze moves beyond the mask, the object is 'attached' to gaze (i.e., the object follows gaze about

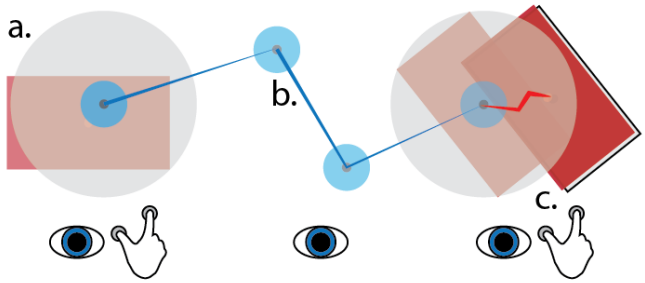

Figure 4. MagicT: (a) Look at object, touch on tablet to select, touch translates and manipulates while the object is within the displayed mask. (b) Gaze is free to move and has no effect on object. (c) Touch translation snaps the object to the gaze location, touch performs final translation and manipulation.

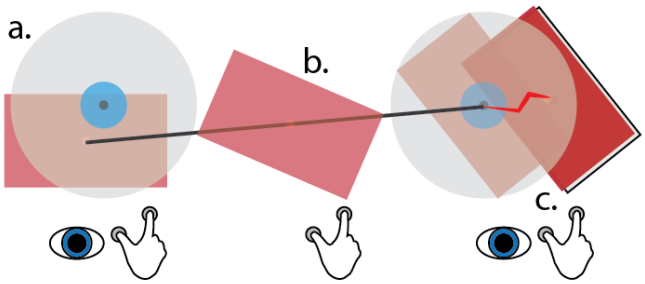

Figure 5. GazeGuidedTouchT: (a) Look at object, touch on tablet to select, touch translates and manipulates freely while the object is within the displayed mask. (b) Touch translates toward gaze, object is snapped to a line connecting to gaze location, constraining translation to shortest path. (c) Object reaches gaze and becomes freely moveable again for final positioning and manipulation.

the display). The object's position updates instantaneously contingent on the user's point-of-regard. (c) At the target destination, dragging with touch detaches the object from gaze control for final touch-based adjustment.

Although though this technique begins and ends with integral touch input, switching to gaze for longer translation changes the input structure to separable. Without experimentation it is unclear if this will be true as all dimensions can technically be controlled concurrently. Our initial thoughts on this technique are that performance may be improved regardless of the input structure mismatch as the eyes translate quickly.

MagicT

This technique is inspired by Zhai et al's MAGIC Pointing [26]. Its input structure aims to maintain integral multi-touch RST while instantly changing the location of the object as gaze moves from the origin of selection to the target.

Figure 4(a) shows a mask after the object has been selected, the mask follows gaze. To perform touch translation, gaze and the object must both reside in this mask. (b) the user looks toward the target and the mask follows. (c) translating with touch snaps the object to the centre of the mask. To reiterate, once touch translation is detected, the object is instantly warped to the location of gaze, but does not follow gaze continuously. Warping does not occur when the object is within the mask surrounding the user's point-of-regard. Final

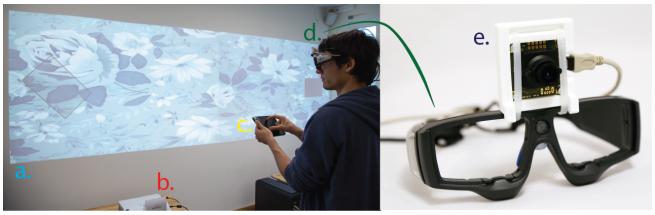

Figure 6. System Hardware Setup: (a) Projected Display. (b) Short-throw projectors. (c) Tablet device. (d) SMI Eye-tracking Glasses. (e) Scene-camera replacement.

translation and manipulation can then be performed before dropping the object.

We believe that the input structure of this technique will demonstrate some level of integrality. Although gaze denotes the warp position, touch-based translation still controls overall movement. This technique also offers a potential advantage over **GazeT** as the eyes do not affect translation unless explicitly triggered, mitigating midas touch issues [10].

GazeGuidedTouchT

Our final technique aims to maintain the integrality of touch throughout interaction while gaze aids in targeting by tying the object to a fixed path between the object's location and gaze position. Using this strategy, the object takes the shortest path to the target (i.e., the object follows a straight line between its current location and the user's point-of-regard).

In Figure 5(a) the object is selected, and a mask appears. Within this mask, the object may be moved along any path. The mask is contingent on gaze. (b) The user looks at the target, causing the mask to follow. At this stage a fixed straight-line path between the object centre and gaze is plotted. If the object moves toward gaze, it is snapped to this line. If the object is moved opposite to this trajectory, it is freed from the line. (c) Touch translates the object until it reaches the mask around gaze, the object is then freed for final positioning.

This method is interesting as it is multimodal but could retain the integrality of touch input. By constraining translation, we aim to reduce task complexity, allowing the user to focus more on rotation and scaling during transit. As described earlier, this technique uses a gaze-contingent CDGain to speed up translation over large distances.

SYSTEM IMPLEMENTATION

We developed an eye-tracking setup that allowed for remote gaze and touch interaction on a large wall display. This is our primary motivation and application context, where objects are translated and manipulated on large screen from a distance. Figure 6 shows our setup.

Our experimental setup was centred around a large projected wall display (see Figure 6(a)). It consisted of two short-throw LCD projectors mounted alongside each other 80cm from the floor (see Figure 6(b)). Each projector's native resolution was 1280 x 800 (WXGA 16:10), projecting a display 2m x 1.24m in size. The two projectors combined had a total resolution of 2560 x 800 and a physical size of 4m x 1.24m. From a centred 2m distance the full display was 90° x 34.5° of visual angle. Both projectors were attached to a PC which ran our eye-tracking/computer vision software and study application.

We used a pair of SMI Eye-tracking Glasses (ETG) (binocular eye-tracker) alongside the SMI SDK to capture realtime gaze data (30Hz), and perform calibration (see Figure 6(d)). These data were fed into our display detection system written in C++ with OpenCV 2.4.9 to be remapped to our projected display. To detect displays, we acquired images of a user's field-of-view via a scene camera. The scene camera included with the ETG system had a significant latency which confounded our techniques. Instead we affixed an 87 fps uEye-1221LE-C USB camera to the top of the ETG on a 3D printed mount (see Figure 6(e)). This allowed us to control exposure and reduce latency. Although the ETG camera was replaced, we were still able to use SMI's 3-point calibration procedure.

To map gaze to our projected display we needed to first detect the display within the scene camera images, we then computed a homography to transform gaze data from scene coordinates to display coordinates. As only portions of the display were visible in scene camera images, we adopted the same approach used by Baur et al. in their Virtual Projection system [2]. Our system took live screen-grabs (15fps) of our display's contents, along with scene camera images. We then used OpenCV's SURF to compute feature-descriptions and match key points between the two images. From these key points we could compute a homography, which in turn was applied to incoming gaze data. For SURF to be successful, the background of our display had to be 'feature rich'. We used a floral design with no repeating patterns that allowed for reliable tracking across the full span of the display. SURF descriptor calculations were performed via OpenCV's CUDA GPU interface, enabling a detection frame rate of 25fps.

Finally our user interface consisted of two parts, one on the projected display, written in WPF and the other on an Apple iPad Mini (1st gen), written in Objective-C (see Figure 6(c)). The iPad captured touch input and recognised pan, pinch, and rotation gestures. Touch data were transmitted via UDP and WiFi to our main PC and study application. The study application was responsible for the study logic and logging of gaze, touches, and experimental measures.

Accuracy & Precision. Our system accuracy was tested using 15 points spread equidistant in a 5 x 3 layout spanning the projected display. On average we found an accuracy of 2.15° however, we found that the left and right extremes of the display showed varying accuracies 2.18° (left) and 4.14° (right). The average precision was found to be 1.62°.

STUDY: INTEGRATING GAZE & RST

This study compares our four techniques with the following questions in mind: (Q1) *Can gaze offer any performance benefits during remote RST tasks?* (Q2) *Do translation distance, rotation, and scale factors affect performance?* (Q3) *Do different configurations of gaze and RST affect integrality?*

Design. To answer Q1 we compared overall completion time and time-to-target performance. We incorporate Q2 by including three additional factors: distance, rotation, and scale, which vary the complexity of each measure. To answer Q3

we record time series of each trial that include object position, scale, rotation, and touch input. These data are used to calculate each technique's integrality ratio. In all, we followed a 4x3x3x3 repeated-measures design (4 techniques x 3 distances x 3 rotations x 3 scales).

First, we defined three distances based on visual angle relative to the user: D40 (40°-685px), D60 (60°-1332px), and D80 (80°-2148px). These distances encompass the full span of the display. We're interested to understand how varying distances might affect performance, particularly given that gaze can move quickly over large distances. Second, we chose three rotations that users would perform during tasks: R0 (0°) R45 (45°), and R90 (90°). These values allowed us to (a) record trials with no rotation at all for later comparison with trials that had rotation and (b) observe how varying levels of rotation would affect performance. Finally, we chose three scale factors S08 (0.8), S1 (1.0), and S12 (1.2), these values are the factor by which targets are sized with respect to a set default. S1 corresponds to no scaling at all, while S08 and S12 correspond to scaling down and up respectively.

The above task factors correspond to the location, size, and orientation of targets in each trial. As a result of our accuracy test, target dimensions were set to 6° x 9° (135px x 200px) at the smallest scale factor (S08), to ensure easy selection and re-selection. Objects were always 11° x 7° (250px x 170px) at the start of each trial and scaled accordingly by the user to match the target. This was also the default target size (S1).

All gaze-based conditions used a mask to delimit states. This mask was sized to be the diameter of the largest possible target size ($250 * 1.2 = 300$px). This ensured participants could interact without error, within the bounds of a target.

To experience every task factor combination, 27 trials had to be completed. We split our study in to 3 blocks of 27 trials per technique. The first block (T) allowed participants to gain aptitude before moving on to two recorded blocks (A & B).

Participants. Sixteen participants (10M 6F) aged between 20 and 43 (M=28.06 SD=5.17) volunteered for our experiment. Participants could not wear glasses due to the small size of our head-worn eye-tracker. If necessary, participants wore contact lenses (3/16). Two participants were left-handed and 14 were right-handed. All participants owned a smartphone and 9/16 owned a multitouch tablet. Eight participants had used eye-tracking more than three times, one three times, one twice, and four once. Two had never used eye-tracking.

Procedure. Participants were first asked to complete a demographic questionnaire then asked to stand on a mark 2m away, parallel with and centred on the projected display. Participants were fitted with the eye-tracker and asked to stand still while a 3-point calibration was completed. We found that calibration drifted over time, to combat this, the procedure was repeated when inaccuracies were discovered. After calibration, participants were given the tablet to hold, ready to begin the first set of tasks.

Conditions were counterbalanced using a balanced latin square, resulting in four unique orders. Upon starting, an object and target would appear on either side of the projected display. These were centred vertically and only varied in their horizontal distance inline with our distance factor. Between participants we altered the direction that objects were moved i.e., left-to-right and right-to left, this was not for factorial analysis but to ensure variance of direction in the dataset.

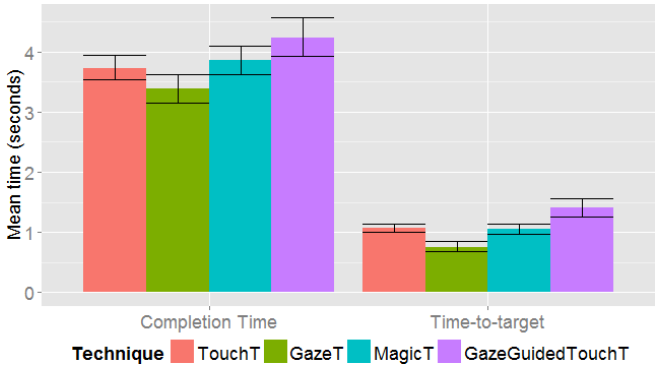

Figure 7. (left) Mean completion time for each technique, (right) Mean time-to-target for each technique. Bars show 95% confidence intervals.

Objects were initially red in colour. Once an object had been selected, participants moved, sized, and orientated it to fit inside the target. Once the object was suitably fit inside the target it would change colour to green, informing the participant to drop, then the next trial would load. We imposed this feedback to encourage participants not to dwell on fitting the object exactly but instead to complete each trial as quickly as possible. This was explicitly explained to each participant.

After completing all three blocks, participants were asked to rate their agreement with statements on usability (9 statements) and aspects of the task (7 statements) on 7-point Likert scales. Participants were also asked to explain what they liked/disliked about a technique, and if any aspects seemed intuitive/easy or confusing/difficult. Finally at the end of the experiment, participants were asked to rank each technique from 1 to 4, 1 being the best, 4 being the worst, and to provide any additional feedback they might have.

RESULTS

In total, participants performed $3 * 27 = 81$ trials of which we analyse the last 27 (Block B). Analysis of these data aimed to ensure our results were representative of the highest aptitude achieved by participants. This equates to 16 repetitions of each factor combination in aggregate. We cover metrics that may provide insight for the questions outlined in our design.

Performance (Q1)

We compared aggregate mean trial completion times between techniques in a one-way repeated-measures ANOVA. Significance was found ($F(3,45)=2.89, p=.045$), but as we use bonferroni correction to reduce false positives, post-hoc tests could not reveal the source of significance. Figure 7(left) summarises the overall means.

Each of our techniques differed in design between selection and target entry. We therefore calculated the mean time taken between these stages i.e., time-to-target. These data

	TouchT F(1.46, 21.9)=30.4,p<.001		GazeT F(1.46,21.9)=4.08,p=.042		MagicT F(1.28,19.2)=11.6,p=.001		GazeGuidedTouchT F(1.22,18.3)=4.9,p=.034	
	D40	**D60**	**D40**	**D60**	**D40**	**D60**	**D40**	**D60**
D60	**p<.001**	-	p=1	-	p=.633	-	p=.160	-
D80	**p<.001**	**p<.005**	p=.139	**p=.044**	**p<.008**	**p=.010**	**p<.001**	p=1

Table 1. Summarised analysis of individual technique performance over distance levels. Significant values are in bold.

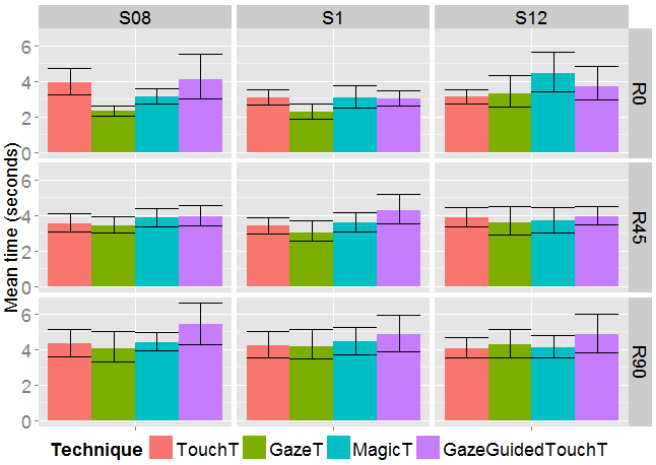

Figure 8. Mean completion time for each technique for all scaling and rotation trials (aggregate distance) with 95% confidence intervals.

are summarised in Figure 7(right). Our aim was to understand how differing transit methods may have affected performance given inconclusive results in overall completion time. An ANOVA found significance (F(3,45)=8.68, p=.001). Post-hoc tests revealed that users were able to reach targets faster with **GazeT** than with **TouchT** (p=.044) and **GazeGuidedTouchT** (p<.001). We expected faster transit with **GazeT** as gaze is used to perform the bulk of object transit. Using **TouchT** and **MagicT** users showed similar target reach times. No further significances were found. Time-to-target series are used again later to provide answers for Q3.

Distance, Rotation, and Scaling (Q2)

We chose to partition the analysis of these factors. We first examine the influence of distance, and then analyse rotation and scaling in all combinations. Distance is isolated to understand if gaze provided a performance boost over particular distances, regardless of rotation and scaling levels. Techniques are analysed individually across their full time series.

Using four one-way repeated-measures ANOVAs, we compared mean completion time over distance levels D40, D60, and D80 within each technique. Each ANOVA revealed significance. Figure 9 summarises the means used for this analysis and Table 1 shows the ANOVA and paired t-test (bonferroni corrected) results. For **TouchT**, performance decreased as distance increased over all levels. In **GazeT**, completion time was slower over D80 compared with D60. With **MagicT**, D80 trials were slower than D40 and D60 trials. Finally with **GazeGuidedTouchT**, D80 trials were slower than D40 trials. This is a little surprising, particularly for **GazeT** and **MagicT** where gaze translation could be close to instantaneous.

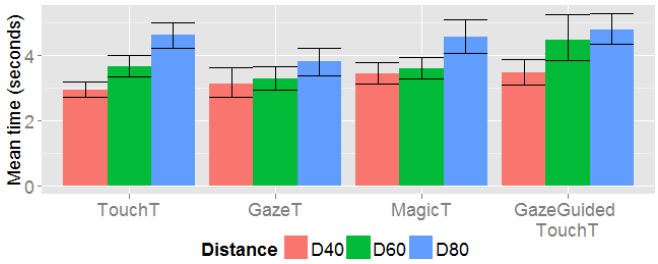

Figure 9. Mean completion time for each technique over distances (aggregate rotation and scaling) with 95% confidence intervals.

For each technique we aimed to understand how trials with no rotation or scaling compared against other level combinations. This was accomplished through four one-way repeated-measures ANOVAs. Each test corresponded to one technique and compared mean completion times for the following factor pairs: R0+S1 (i.e., no rotation or scaling), R45+S1, R90+S1, R0+S08, R45+S08, R45+S08, R0+S12, R45+S12, R45+S12. Significance was found for **GazeT** (F(2.88, 43.2)=6.33,p=.001). Compared to levels of R0+S1 users were significantly slower with R45+S1 (p=.034), R90+S1 (p=.007), R45+S08 (p=.002), R45+S12 (p=.033), R90+S12 (p=.008). No other significances were found.

During observation we noticed that participants had issue with large rotations, turning the wrist to awkward positions instead of "clutching" to compensate. This prompted a short analysis of clutching behaviour. We define clutching as the number of times a user reselected an object not counting the original selection. Table 2 summarises the mean clutches per trial. A one-way repeated measures ANOVA revealed no significant differences between our techniques. In addition we compared techniques over all levels of distance, rotation, and scaling factors. This resulted in a total of nine ANOVAs. Two tests showed significance. Over distances of D80 (F(1.98,29.7)=3.89,p=.032) we found significantly less clutching with **GazeT** (M=3.62 SD=4.86) than with **TouchT** (M=8.06 SD=5.96) (p<.001). Over rotation levels, significance was found when no rotation was required (i.e., R0) (F(3,45)=3.43,p=.036). Post-hoc tests again found that **GazeT** (M=1.06 SD=2.14) required significantly less clutching than **TouchT** (M=3.75 SD=2.96) (p=.009). No differences were found between techniques over scaling levels.

Integrality (Q3)

Here we only use data recorded between selection and the object entering the target bounds (time-to-target series). This ensures we gain a clear picture of users' integrality during transit, where our techniques differ.

First, each technique is designed to utilise a different amount of touch based translation, we measured the mean amount of translation (in pixels) that participants performed during transit. This quantitatively informs us if participants were interacting as intended with our techniques.

Table 2 summarises the mean touch-based translation used in each technique. A one-way repeated-measures ANOVA ($F(3,45)=25.13, p<.001$) revealed significant differences between **TouchT**, **GazeT** and **MagicT** (all $p<.001$). The means tell us that **TouchT** required more touch-based translation than **GazeT** and **MagicT**, which aligns with our design intentions. Additionally **GazeT** and **MagicT** showed less touch-based translation than **GazeGuidedTouchT** ($p<.001$). By design, **GazeT** should only require touch-based translation within the target, but the mean shows 221.3px of translation. Additionally, we might expect a lower mean for **MagicT**.

Secondly, we compare techniques in terms of integrality. As stated by Jacob et al., there are levels integrality and separability and these classes to not form a sharp dichotomy. We use the euclidian to city-block ratio metric to quantify the level of integrality each technique demonstrated, details of the algorithm are summarised below.

The ratio is calculated using the following method taken from Jacob et al. [11]: (1) Resample time series with a 10ms period; (2) Calculate the difference in movement for each DOF and label as active if the difference is greater than the set parameters; (3) If movement only occurs in one dimension, label the sample as "city-block", if two or more dimensions are active, label the sample as "euclidean"; (4) Calculate the ratio of euclidean to city-block samples.

We needed to set values for three thresholds used to classify euclidean and city-block behaviour. These parameters are used to determine if a sufficient amount of translation (P_t), rotation (P_r), or scaling (P_s) has occurred to deem a DOF as 'active'. The analysis we report here holds true through the following range of values $P_t = 0, 1, ..., 3$px, $P_r = 0.0°, 0.1°, ..., 1.0°$, and $P_s = 0, .001, ..., .01$. All reported statistical values pertain to the upper bound of these parameters, and significance is present in the full range.

We compared the mean integrality ratios of all techniques, none of which reached over 1.0 (see Table 2). A value greater than 1.0 would indicate higher euclidean than city-block behaviour. An ANOVA of these means ($F(3,45)=4.53, p=.007$) showed that **GazeT** was significantly less integral than **TouchT** ($p=.032$) and **MagicT** ($p=.025$) while **GazeGuided-TouchT** showed no differences.

Likert Scale Responses

Regarding usability, Friedman tests found significance between conditions for responses on eye fatigue ($\chi^2_3=9.28, p=.026$) and liking techniques ($\chi^2_3=7.91, p=.047$). However, Wilcoxon signed rank tests (Bonferroni corrected $\alpha=.05/6=.0083$) could not reveal the specific significant conditions. No other significant results were found. Mean responses are summarised in Figure 10.

	Clutching (per trial)		*Integrality (ratio)*		*Touch (trans. in px)*	
	M	**SD**	**M**	**SD**	**M**	**SD**
TouchT	.59	.47	.63	.45	483.69	122.98
GazeT	.34	.48	.33	.36	221.3	222.79
MagicT	.5	.56	.67	46	298.67	122.46
GazeGuided-TouchT	.56	.67	.58	.52	526.38	179.50

Table 2. Summarised mean and standard deviation results from our coordination analysis.

For responses regarding task aspects, significance was found for 'moving objects easily' ($\chi^2_3=12, p=0.0073$) and post-hoc tests ($\alpha=.0083$) revealed a significant difference between **GazeT** (M=6.63 SD=0.81) and **GazeGuidedTouchT** (M=5.7 SD=1.20) ($p=.006$), suggesting that participants found it easier to move objects with **GazeT**, where objects were transported via gaze. No other significant results were found. Mean responses are summarised in Figure 11.

Feedback & Observations

For overall preference, participants ranked each technique from best to worst. For best ranking, the frequencies were **GazeT**=9, **MagicT**=3, **TouchT**=2, **GazeGuidedTouchT**=2.

We identified several themes from observation and subjective feedback that highlight interesting usability aspects. Across all techniques users highlighted problems with large rotations, sometimes ending up in awkward positions instead of clutching to compensate.

Generally users found **TouchT** easy to use (P1, P9, P13, P14) and those that had used **GazeGuidedTouchT** previously noted that translation was easier when not tied to a particular trajectory (P4, P12, P16). Only one participant noticed and liked the gaze-contingent CDGain (P6). Several found touch translation laborious (P1, P3, P7, P13) and fatiguing (P2).

Fifteen participants stated that **GazeT** felt fast and noted that the translation of objects was very easy. P6 pointed out that for short distances they would drag the object with touch, following with gaze, as opposed to moving with gaze outright. We believe this strategy may be why we saw an amount of touch-based translation in our quantitative analysis. P13 felt they could not rotate in a concurrent manner as the object moved too quickly, reducing coordination. P2 pointed out that for larger distances they didn't know the orientation of the target; due to the instant translation of the object they could not rotate prior to reaching the target.

Participants noted two strategies that they used with **MagicT**. The first used gaze and touch translation in unison: as gaze moved across the display, touch dragged as well, constantly updating the position of the object in a coarse manner (P1, P2). The second involved selecting the object: moving the eyes to the target, then translating slightly to snap the object into place (P3, P6, P8, P16). This is how we intended users to interact. On occasion, users forgot they had selected an object and went back to the origin to ensure it was selected despite the feedback provided by the technique (P5, P6, P10, P13).

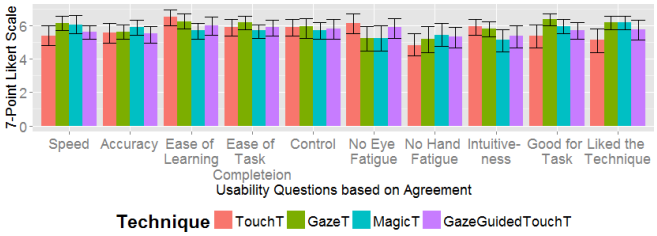

Figure 10. Usability: Mean Likert Scale Responses

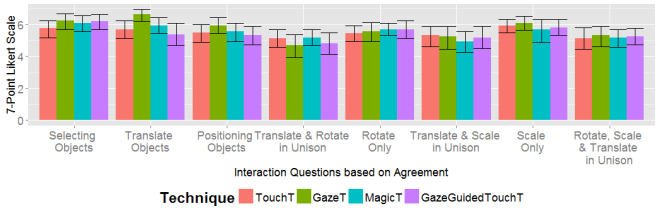

Figure 11. Interaction: Mean Likert Scale Responses

For **GazeGuidedTouchT** only three participants (P1, P5, P13) noted liking the targeting guidance in this technique. From our observations during trials, the guidance line became increasingly ignored as the study progressed. P3, P4, P10, and P15 expressed dislike for this particular aspect, finding the sequence of interaction confusing.

DISCUSSION & CONCLUSION

Overall our techniques were found to be usable by all participants, completing all 27 tasks in block B without error. Despite not finding an overall best performer, we found that **GazeT** offered a speed benefit during transit to a target (Q1). However, using gaze in this configuration came at a cost. Trials with rotation and scaling were significantly slower than those without, which was found to be unique to **GazeT** (Q2). Furthermore **GazeT** was found to be the least integral of all techniques. Although users had the ability to operate all dimensions concurrently, lower integrality confirms that rotation and scaling were constrained during transit, classifying this input structure as separable. We attribute this to the speed of the gaze, leaving little transit time for concurrent manipulation (Q3). However, this confirms that gaze is beneficial for translation-only tasks, aligning with gaze-supported positioning research [18, 19]. The reduced amount of clutching and touch-based translation compared with **TouchT** additionally supports this benefit.

MagicT showed similar time-to-target results as **TouchT**, interestingly showing no boost in performance (Q1). Similarly, there was little difference in integrality ratios or performance over varying task factors (Q2). Subjective feedback pointed out that some users would continually drag during transit to target. The eyes pre-empt the location of interaction, and we would expect users to look directly to the target (as in Zhai et al. [26]), but instead some users pursued objects, counteracting **MagicT**'s mechanics. This phenomenon was not expected and never reported by Zhai et al. however a smaller mask may reduce this effect, unless it pertains to natural user behaviour. In general, all techniques' performance suffered over larger distances, which was not expected for **GazeT** and **MagicT**. It is also possible that head movement over larger distances may have contributed to lowered performance.

We see significantly less touch-based translation compared to **TouchT**, but a similar amount to **GazeT**. So it is clear that **MagicT** and **GazeT** require less touch translation. With **MagicT** we do however see a higher integrality ratio than **GazeT** and a similar ratio to **TouchT**. This means that **MagicT** is able to combine gaze and touch while maintaining similar integrality to touch-only interaction (Q3). As **MagicT** warps objects to gaze, we would expect better performance (as in **GazeT**). It is perhaps possible that the need to snap content to gaze required additional overhead from users.

TouchT behaved as expected. The gaze-contingent CDGain used equalised translation speed compared to other techniques. Although not well favoured, it provides a simple and viable solution while maintaining integrality.

From observation and subjective feedback it is clear that **GazeGuidedTouchT** is not an optimal design for such a task. It was often the case that the object would reach the target ahead of gaze, overshooting, and thus not benefiting from the technique's object guidance mechanic (Q1). This occurred because of the high CDGain applied when objects were not within the low gain mask. Additionally, some users would simply drag the object across the display while continually fixating on it, this caused a low CDGain throughout movement and the behaviour became identical to **TouchT** (Q3). This is confirmed in the touch-based translation means. Observations revealed that object guidance was generally ignored.

A limitation of this study is a lack of repetitions. We may have yielded clearer performance results with a reduced factor set, however we have learned that increasing levels of task complexity can affect both touch-only and gaze-based techniques (Q2). The 4x3x3x3 design aimed to get an overview of how different manipulation combinations would affect tasks, but it is clear now the that the high variability of tasks influences overall performance. Furthermore, users had difficulty with large rotations. Users were mandated to complete tasks as fast as possible, and perhaps this motivation encouraged a lack of clutching. It is possible a 'fear' of dropping objects existed, however it was not reported by users. In addition, our results lack time-course analyses to gauge the distribution of DOF control over time. We consider this for future work.

Design Implications

Despite the above limitations, we believe our findings can inform future gaze-based technique designs, in particular when considering more complex tasks. From the results that we have, we believe that **TouchT**, **GazeT**, and **MagicT** all present viable options with varying strengths and weaknesses.

TouchT allows for full touch control with the additional benefit of gaze-supported remote selection in out-of-reach contexts. It is most suited to integral tasks.

GazeT demonstrates its strength in using gaze for object transit, however users may struggle to perform concurrent oper-

ations, and therefore it is best suited to separable tasks. This may explain Stellmach and Dachselt's need for a mode switch to allow rotation and scaling [18].

MagicT allows gaze to support translation with concurrent rotation and scaling. The benefit over **TouchT** is the reduced amount of touch-translation which may be suited to situations where touch input space is limited i.e., on a smartphone or smartwatch. Additionally, as gaze warps objects, longer distances could yield performance benefits e.g., multi-wall.

CONCLUSION

In summary, we have developed four novel interaction techniques that combine gaze and multitouch RST for interaction with large remote displays. Techniques were designed to answer three: questions. (1) *Can gaze offer any performance benefits during remote RST tasks?* (2) *Do translation distance, rotation, and scale factors affect performance?* (3) *Do different configurations of gaze and RST affect integrality?*

Overall we are yet to see a performance benefit from the inclusion of gaze for remote RST. We have learned that distance, rotation, and scaling factors do not affect performance in integral techniques, but do in separable techniques. Finally, in order to achieve integrality with gaze and multitouch, gaze must be secondary or cascaded in configuration. Using gaze as the primary positioning modality causes difficulty in orchestrating multiple degrees-of-freedom.

REFERENCES

1. Balakrishnan, R., and Hinckley, K. Symmetric bimanual interaction. In *Proc. CHI '00* (2000).
2. Baur, D., Boring, S., and Feiner, S. Virtual projection: Exploring optical projection as a metaphor for multi-device interaction. In *Proc. CHI '12* (2012).
3. Boring, S., Baur, D., Butz, A., Gustafson, S., and Baudisch, P. Touch projector: Mobile interaction through video. In *Proc. CHI '10* (2010).
4. Dachselt, R., and Buchholz, R. Natural throw and tilt interaction between mobile phones and distant displays. In *Proc. CHI EA '09* (2009).
5. Fares, R., Downing, D., and Komogortsev, O. Magic-sense: Dynamic cursor sensitivity-based magic pointing. In *Proc. CHI EA '12* (2012).
6. Fares, R., Fang, S., and Komogortsev, O. Can we beat the mouse with magic? In *Proc. CHI '13* (2013).
7. Garner, W. R. *The processing of information and structure*. Psychology Press, 2014.
8. Grasso, M. A., Ebert, D., and Finin, T. The effect of perceptual structure on multimodal speech recognition interfaces. *organization* (1998).
9. Han, S., Lee, H., Park, J., Chang, W., and Kim, C. Remote interaction for 3d manipulation. In *CHI EA '10* (2010).
10. Jacob, R. J. K. What you look at is what you get: Eye movement-based interaction techniques. In *Proc. CHI '90* (1990).
11. Jacob, R. J. K., Sibert, L. E., McFarlane, D. C., and Mullen, Jr., M. P. Integrality and separability of input devices. *ACM Trans. Comput.-Hum. Interact.* (1994).
12. Keefe, D. F., Gupta, A., Feldman, D., Carlis, J. V., Keefe, S. K., and Griffin, T. J. Scaling up multi-touch selection and querying: Interfaces and applications for combining mobile multi-touch input with large-scale visualization displays. *International Journal of Human-Computer Studies 70*, 10 (2012), 703 – 713.
13. Martinet, A., Casiez, G., and Grisoni, L. The effect of dof separation in 3d manipulation tasks with multi-touch displays. In *Proc. VRST '10* (2010).
14. Masliah, M. R. Quantifying human coordination in hci. In *CHI EA '99* (1999).
15. Nancel, M., Wagner, J., Pietriga, E., Chapuis, O., and Mackay, W. Mid-air pan-and-zoom on wall-sized displays. In *Proc. CHI '11* (2011).
16. Pfeuffer, K., Alexander, J., Chong, M. K., and Gellersen, H. Gaze-touch: Combining gaze with multi-touch for interaction on the same surface. In *Proc. UIST '14* (2014).
17. Sibert, L. E., and Jacob, R. J. K. Evaluation of eye gaze interaction. In *Proc. CHI '00* (2000).
18. Stellmach, S., and Dachselt, R. Still looking: Investigating seamless gaze-supported selection, positioning, and manipulation of distant targets. In *Proc. CHI '13* (2013).
19. Turner, J., Alexander, J., Bulling, A., Schmidt, D., and Gellersen, H. Eye pull, eye push: Moving objects between large screens and personal devices with gaze and touch. In *Proc. INTERACT '13* (2013).
20. Turner, J., Bulling, A., Alexander, J., and Gellersen, H. Cross-device gaze-supported point-to-point content transfer. In *Proc. ETRA '14* (2014).
21. Veit, M., Capobianco, A., and Bechmann, D. Influence of degrees of freedom's manipulation on performances during orientation tasks in virtual reality environments. In *Proc. VRST '09* (2009).
22. Vogel, D., and Balakrishnan, R. Distant freehand pointing and clicking on very large, high resolution displays. In *Proc. UIST '05* (2005).
23. Wang, Y., MacKenzie, C. L., Summers, V. A., and Booth, K. S. The structure of object transportation and orientation in human-computer interaction. In *Proc. CHI '98* (1998).
24. Yoo, B., Han, J.-J., Choi, C., Yi, K., Suh, S., Park, D., and Kim, C. 3d user interface combining gaze and hand gestures for large-scale display. In *CHI EA '10* (2010).
25. Zhai, S., and Milgram, P. Quantifying coordination in multiple dof movement and its application to evaluating 6 dof input devices. In *Proc. CHI '98* (1998).
26. Zhai, S., Morimoto, C., and Ihde, S. Manual and gaze input cascaded (magic) pointing. In *Proc. CHI '99* (1999).